Protein databases

Henrik Nielsen

Protein databases, historical background

Swiss-Prot, http://www.expasy.org/sprot/
Established in 1986 in Swizerland
ExPASy (Expert Protein Analysis System)
Swiss Institute of Bioinformatics (SIB) and European Bioinformatics Institute (EBI)

PIR, http://pir.georgetown.edu/
Established in 1984
National Biomedical Research Foundation, Georgetown University, USA

In 2002 merged into:
UniProt, http://www.uniprot.org/
A collaboration between SIB, EBI and Georgetown University.

UniProt

UniProt Knowledgebase (UniProtKB)
UniProt Reference Clusters (UniRef)
UniProt Archive (UniParc)

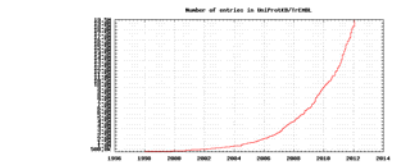
UniProt Knowledgebase Release 2012_05 (May 16, 2012) consists of:

UniProtKB/Swiss-Prot: Annotated manually (*curated*)
536,029 entries, 190,235,160 amino acids

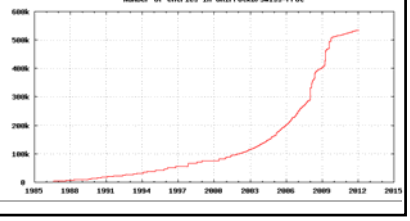
UniProtKB/TrEMBL: Computer annotated (automatically translated from nucleotide databases)
22,128,511 entries, 7,226,807,757 amino acids

Growth of UniProt

TrEMBL



Swiss-Prot



Content of UniProt Knowledgebase

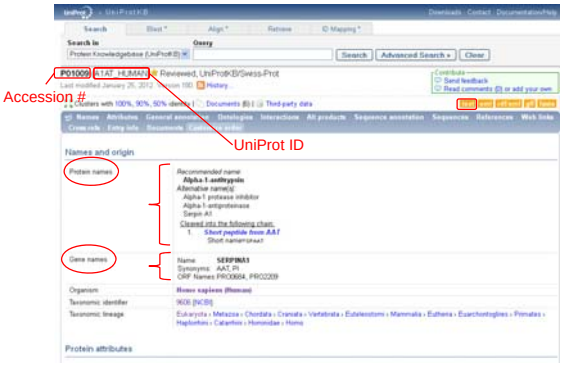
- Amino acid sequences
- Functional and structural annotations
  - Function / activity
  - Secondary structure
  - Subcellular location
  - Mutations, phenotypes
  - Post-translational modifications
- Origin
  - organism: Species, subspecies; classification
  - tissue
- References
- Cross references

Amino acid sequences

From where do you get amino acid sequences?

- Translation of nucleotide sequences (GenBank/EMBL/DDBJ)
- Direct amino acid sequencing: *Edman degradation*
- Mass spectrometry
- 3D-structures

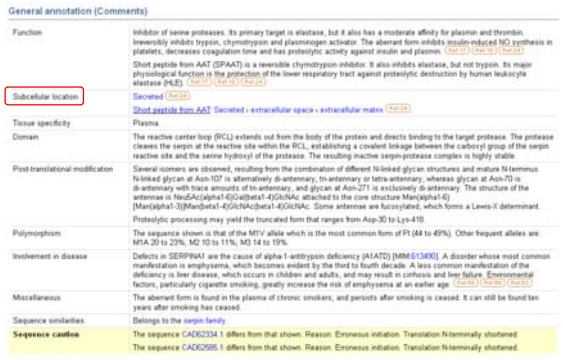
UniProt entry, formatted view



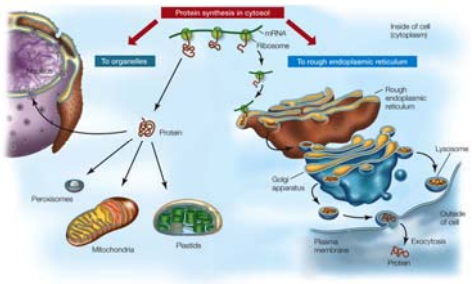
UniProt entry, text view (flat file)

```
ID   A1AT_HUMAN              Reviewed;         418 AA.
AC   P01009; A6PX14; B2RDQ8; Q0PVP5; Q13672; Q53XB8; Q5UUM1; Q7M4R2;
AC   Q86U18; Q86U19; Q96BF9; Q96ES1; Q9P1P0; Q9UCE6; Q9UCM3;
DT   21-JUL-1986, integrated into UniProtKB/Swiss-Prot.
DT   01-OCT-1996, sequence version 3.
DT   25-JAN-2012, entry version 180.
DE   RecName: Full=Alpha-1-antitrypsin;
DE   AltName: Full=Alpha-1 protease inhibitor;
DE   AltName: Full=Alpha-1-antiproteinase;
DE   AltName: Full=Serpin A1;
DE   Contains:
DE     RecName: Full=Short peptide from AAT;
DE              Short=SPAAT;
DE   Flags: Precursor;
GN   Name=SERPINA1; Synonyms=AAT, PI; ORFNames=PRO0684, PRO2209;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC   Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1).
RX   MEDLINE=84107980; PubMed=6319097;
RA   Bollen A., Herzog A., Cravador A., Herion P., Chuchana P.,
RA   van der Straten A., Loriau R., Jacobs P., van Elsen A.;
RT   "Cloning and expression in Escherichia coli of full-length
RT   complementary DNA coding for human alpha 1-antitrypsin.";
RL   DNA 2:255-264(1983).
...
```

General annotation (Comments)



Protein sorting in eukaryotes



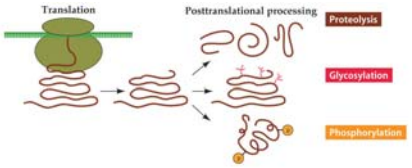
Different proteins belong to different compartments of the cell – and some belong *outside* the cell

General annotation (Comments)

General annotation (Comments)

Function	Inhibitor of serine proteases. Its primary target is elastase, but it also has a moderate affinity for plasmin and thrombin. Irreversibly inhibits trypsin, chymotrypsin and plasminogen activator. The aberrant form inhibits insulin-induced NO synthesis in platelets, decreases coagulation time and has proteolytic activity against insulin and plasmin. [1] [2] [3] [4] [5] [6] [7] Short peptide from AAT (SPAAT) is a reversible chymotrypsin inhibitor. It also inhibits elastase, but not trypsin. Its major physiological function is the protection of the lower respiratory tract against proteolytic destruction by human leukocyte elastase (HLE). [8] [9] [10] [11] [12]
Subcellular location	Secreted [13] [14] Short peptide from AAT: Secreted › extracellular space › extracellular matrix [14] [15]
Tissue specificity	Plasma.
Domain	The reactive center loop (RCL) extends out from the body of the protein and directs binding to the target protease. The protease cleaves the serpin at the reactive site within the RCL, establishing a covalent linkage between the carboxyl group of the serpin reactive site and the serine hydroxyl of the protease. The resulting inactive serpin-protease complex is highly stable.
Post-translational modification	Several isomers are observed, resulting from the combination of different N-linked glycan structures and mature N-terminus. N-linked glycan at Asn-107 is alternatively di-antennary, tri-antennary or tetra-antennary, whereas glycan at Asn-70 is di-antennary with trace amounts of tri-antennary, and glycan at Asn-271 is exclusively di-antennary. The structure of the antennae is Neu5Ac(alpha2-6)Gal(beta1-4)GlcNAc attached to the core structure Man(alpha1-6)[Man(alpha1-3)]Man(beta1-4)GlcNAc(beta1-4)GlcNAc. Some antennae are fucosylated, which forms a Lewis-X determinant. Proteolytic processing may yield the truncated form that ranges from Asp-30 to Lys-418.
Polymorphism	The sequence shown is that of the M1V allele which is the most common form of PI (44 to 49%). Other frequent alleles are: M1A 20 to 23%, M2 10 to 11%, M3 14 to 19%.
Involvement in disease	Defects in SERPINA1 are the cause of alpha-1-antitrypsin deficiency (A1ATD) [MIM:613490]. A disorder whose most common manifestation is emphysema, which becomes evident by the third to fourth decade. A less common manifestation of the deficiency is liver disease, which occurs in children and adults, and may result in cirrhosis and liver failure. Environmental factors, particularly cigarette smoking, greatly increase the risk of emphysema at an earlier age. [16] [17] [18] [19] [20]
Miscellaneous	The aberrant form is found in the plasma of chronic smokers, and persists after smoking is ceased. It can still be found ten years after smoking has ceased.
Sequence similarities	Belongs to the serpin family.
Sequence caution	The sequence CAD62334.1 differs from that shown. Reason: Erroneous initiation. Translation N-terminally shortened. The sequence CAD62585.1 differs from that shown. Reason: Erroneous initiation. Translation N-terminally shortened.

Post-translational modifications



Many proteins need to be *modified* after their synthesis to become active

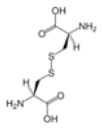
Proteolysis: cleavage of *signal peptides*, *propeptides* or *initiator methionine*

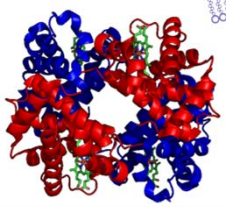
Glycosylation: Especially relevant on the cell *surface*. Also plays a role in sorting of proteins to *lysosomes*

Phosphorylation: Often *reversible*. Regulates the *activity* of many enzymes

More post-translational modifications

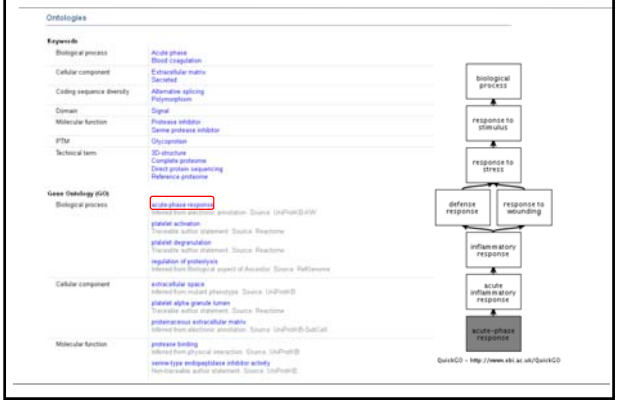
- Lipid anchors
  - (e.g. GPI anchors)
- Disulfide bonds





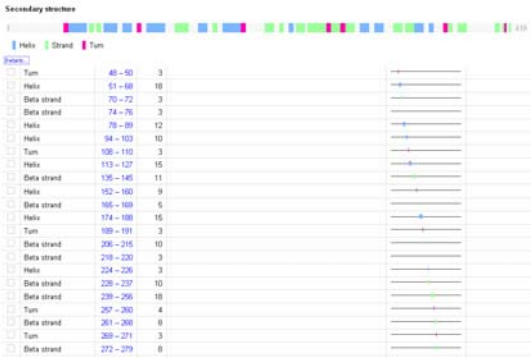
- Prosthetic groups
  - (*e.g.* metal ions)

General annotation (Ontologies)



Sequence annotation (Feature Table)

Secondary structure (Feature Table)



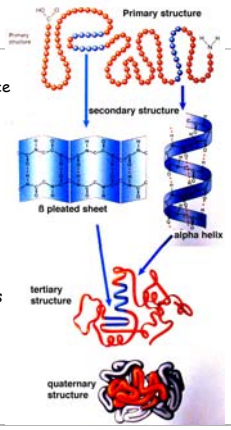
Protein structure

Primary structure: Amino acid sequence

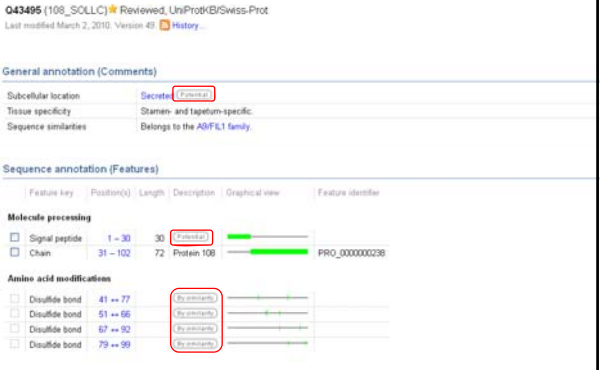
Secondary structure:
"Backbone" hydrogen bonding
Alpha helix / Beta sheet / Turn

Tertiary structure: Fold, 3D coordinates

Quaternary structure: subunits



Evidence (Comments, Feature Table)

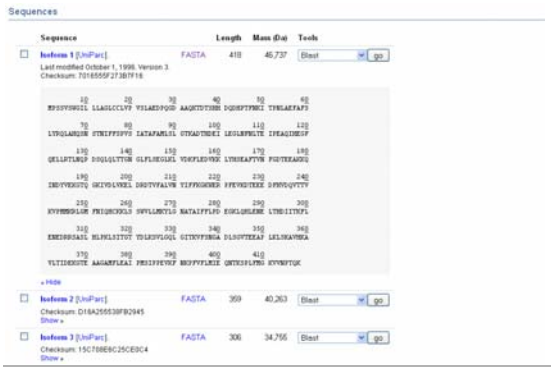


Evidence/Confidence types

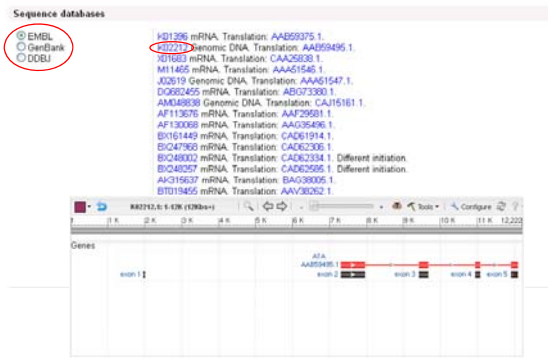
3 types of *non-experimental qualifiers* in Sequence annotation and General comment:

- *Potential:* Predicted using sequence analysis
- *Probable:* Uncertain experimental evidence
- *By similarity:* Predicted using sequence similarity

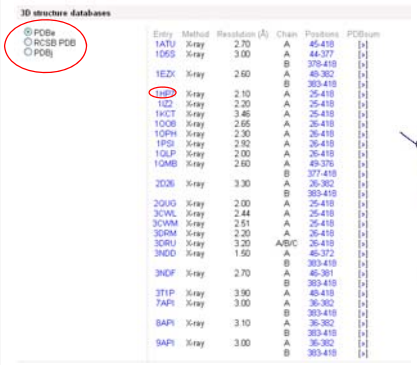
UniProt entry, sequence(s)



Cross-references, nucleotide sequences



Cross-references, 3D structure



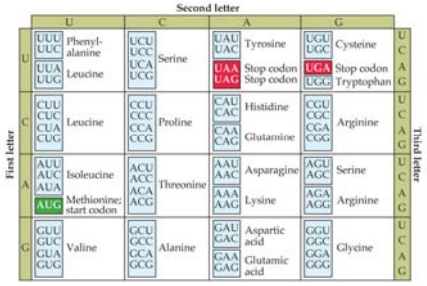
Cross-references

Other databases linked from UniProt
(there are ~100 in total):

- Nucleotide sequences
- 3D structure
- Protein-protein interactions
- Enzymatic activities and pathways
- Gene expression (microarrays and 2D-PAGE)
- Ontologies
- Families and domains
- Organism specific databases

Translation and Reading Frames

The genetic code



- Degenerate (*redundant*) but not ambiguous
- *Almost* universal (deviations found in mitochondria)

Reading Frames 1

A piece of an mRNA-strand:

```
5'                                                    3'
augcccaagcugaauagcguagagggguuuucaucauuugaggacgauguauaa
```

can be divided into triplets (*codons*) in three ways:

```
 1 aug ccc aag cug aau agc gua gag ggg uuu uca uca uuu gag gac gau gua uaa
   M   P   K   L   N   S   V   E   G   F   S   S   F   E   D   D   V   *
 2  ugc cca agc uga aua gcg uag agg ggu uuu cau cau uug agg acg aug uau
    C   P   S   *   I   A   *   R   G   F   H   H   L   R   T   M   Y
 3   gcc caa gcu gaa uag cgu aga ggg guu uuc auc auu uga gga cga ugu aua
     A   Q   A   E   *   R   R   G   V   F   I   I   *   G   R   C   I
```

Each possible set of triplets is called a *reading frame.*

Reading Frames 2

Since there are two strands in DNA, there are *six* possible reading frames in a piece of DNA (three in each direction):

```
3    A  Q  A  E  *  R  R  G  V  F  I  I  *  G  R  C  I
2   C  P  S  *  I  A  *  R  G  F  H  H  L  R  T  M  Y
1  M  P  K  L  N  S  V  E  G  F  S  S  F  E  D  D  V  *
5' ATGCCCAAGCTGAATAGCGTAGAGGGGTTTTCATCATTTGAGGACGATGTATAA 3'
3' TACGGGTTCGACTTATCGCATCTCCCCAAAAGTAGTAAACTCCTGCTACATATT 5'
    H  G  L  Q  I  A  Y  L  P  K  *  *  K  L  V  I  Y  L  -1
     G  L  S  F  L  T  S  P  N  E  D  N  S  S  S  T  Y  -2
    A  W  A  S  Y  R  L  P  T  K  M  M  Q  P  R  H  I   -3
```

A reading frame from a start codon to the first stop codon is called an *open* reading frame (underlined above).